Appendix A

Variable Description and Group Differences

Caio Brighenti

Variable Description

This section briefly lists and describes each variable available in the dataset of text features for real and fake articles. Each variable is a metric quantifying the semantic properties of the text body.

Complexity Metrics

These features are intended to capture the complexity of the text. Some of these are direct textual complexity metrics calculated using the `quanteda` (Benoit et al. 2018) package in R. Others describe the mean, standard deviation, and interquatile range of text parse tree depths calculated using the Stanford Core NLP (Manning et al. 2014) toolset, which constructs consituency trees for each sentence in the text body. There are other miscelleanouscomplexity metrics described below.

- `mu_sentence` - Mean number of sentences
- `mu_verb_phrase` - Mean depth of verb-phrase trees
- `mu_noun_phrase` - Mean depth of noun-phrase trees
- `sd_sentence` - Standard deviation of number of sentences
- `sd_verb_phrase` - Standard deviation of depth of verb-phrase trees
- `sd_noun_phrase` - Standard deviation of depth of noun-phrase trees
- `iqr_sentence` - Interquantile range of number of sentences
- `iqr_verb_phrase` - Interquantile range of depth of verb-phrase trees
- `iqr_noun_phrase` - Interquantile range of depth of verb-phrase trees
- `num_verb_phrase` - Number of verb-phrase trees
- `swc` - Mean sentence word count
- `wlen` - Mean word length
- `types` - Number of unique words
- `tokens` - number of total words
- `TTR` - Type-token ration
- `FOG` - Gunning's Fog Index
- `SMOG` - Simple Measure of Gobbledygook
- `FK` - Flesch-Kincaid Readability Score
- `CL` - Coleman-Liau Index
- `ARI` - Automated Readability Index

LIWC Dictionary

The variables in this section come from the Linguistic Inquiry and Word Count (Pennebaker et al. 2015) dictionary tool. At a high level, LIWC reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. The variables are broken up into conceptual groups.

`summary` - Summary variables

- `WC` - Word count
- `Analytic` - Words reflecting formal, logical, and hierarchical thinking
- `Clout` - Words suggesting author is speaking from a position of authority

- `Authentic` - Words associated with a more honest, personal, and disclosing text
- `Tone` - Words associated with positive, upbeat style
- `WPS` - Words per sentence
- `Sixltr` - Number of six+ letter words
- `Dic` -

`function` - Function words

- `function` - Function words
- `pronoun` - Pronouns
- `ppron` - Personal pronouns
- `i` - 1st person singular
- `we` - 1st person plural
- `you` - 2nd person
- `shehe` - 3rd person singular
- `they` - 3rd person plural
- `ipron` - Impersonal pronoun
- `article` - Articles
- `prep` - Prepositions
- `auxverb` - Auxiliary verbs
- `adverb` - Common adverbs
- `conj` - Conjuctions
- `negate` - Negations

`othergram` - Other grammar

- `verb` - Regular verbs
- `adj` - Adjectives
- `compare` - Comparatives
- `interrog` - Interrogatives
- `number` - Numbers
- `quant` - Quantifiers

`affect` - Affect words

- `affect` - Affect words
- `posemo` - Positive emotions
- `negemo` - Negative emotions
- `anx` - Anxiety
- `anger` - Anger
- `sad` - Sad

`social`

- `social` - Social words
- `family` - Family
- `friend` - Friends
- `female` - Female referents
- `male` - Male referents

`cogproc` - Cognitive processes

- `cogproc` - Cognitive processes

- `insight` - Insight
- `cause` - Cause
- `discrep` - Discrepancies
- `tentat` - Tentativeness
- `certain` - Certainty
- `differ` - Differentiation

`percept` - Perceptual processes

- `percept` - Perceptual processes
- `see` - Seeing
- `hear` - Hearing
- `feel` - Feeling

`bio` - Biological processes

- `bio` - Biological processes
- `body` - Body
- `health` - Health/illness
- `sexual` - Sexuality
- `ingest` - Ingesting

`drives` - Core drives and needs

- `drives` - Core drives
- `affiliation` - Affiliation
- `achieve` - Achievement
- `power` - Power
- `reward` - Reward focus
- `risk` - Risk/prevention focus

`timeorient` - Time orientation

- `focuspast` - Past focus
- `focuspresent` - Present focus
- `focusfuture` - Future focus

`relativ` - Relativity

- `relativ` - Relativity
- `motion` - Motion
- `space` - Space
- `time` - Time

`personc` - Personal Concerns

- `work` - Work
- `leisure` - Leisure
- `home` - Home
- `money` - Money
- `relig` - Religion
- `death` - Death

`informal` - Informal speech

- `informal` - Informal speech
- `swear` - Swear words
- `netspeak` - Netspeak
- `assent` - Assent
- `nonflu` - Nonfluencies
- `filler` - Fillers

`punc` - Punctuation

- `AllPunc` - All punctuation
- `Period` - Periods
- `Comma` - Commas
- `Colon` - Colons
- `SemiC` - Semicolons
- `QMark` - Question marks
- `Exclam` - Exclamation marks
- `Dash` - Dashes
- `Quote` - Quotes
- `Apostro` - Apostrophes
- `Parenth` - Parentheses (pairs)
- `OtherP` - Other punctuation

Parts-Of-Speech Tagging

These variables count the frequency of categories of parts-of-speech (verb, nouns, etc), created using the Stanford CoreNLP (Manning et al. 2014) parts-of-speech tagger. For further detail, see this document detailing each category.

- `CC` - Coordinating conjunctions
- `CD` - Cardinal numeral
- `DT` - Determiner
- `EX` - Existential
- `FW` - Foreign word
- `IN` - Preposition or subordinating conjunction
- `JJ` - Ordinal number
- `JJR` - Comparative adjective
- `JJS` - Superlative adjective
- `LS` - List item marker
- `MD` - Model verb
- `NN` - Noun, singular or mass
- `NNS` - Plural noun
- `NNP` - Singular proper noun
- `NNPS` - Plural proper noun
- `PDT` - Predeterminer
- `POS` - Possessive ending
- `PRP` - Personal pronoun
- `PRP$` - Possessive pronoun
- `RB` - Adverb
- `RBR` - Comparative adverb
- `RBS` - Superlative adverb
- `RP` - Particle
- `SYM` - Symbol

- `TO` - To
- `UH` - Exclamation/interjection
- `VB` - Verb, base form
- `VBD` - Past tense verb
- `VBG` - Present participle
- `VBN` - Past participle
- `VBP` - Present tense verb, other than 3rd person singular
- `VBZ` - Present tense verb, 3rd person singular
- `WDT` - Wh-determiner
- `WP` - Wh-pronoun
- `WP$` - Possessive wh-pronoun
- `WRB` - Wh-adverb

Named Entity Recognition

This includes only a single variable: the number of named entities in the text, counted by the Stanford CoreNLP (Manning et al. 2014) Named Entity Recognition tool.

- `NER` - Number of named entities

Outlier Removal

The dataset of labels (true/fake), article title, and article text body were gathered using the JSON scraper included by the authors of the FakeNewsNet dataset (Shu et al. 2018). However, given that web pages often change structure, move to a different address, or are removed from the internet entirely, many of the entries have changed since the dataset was initially compiled, and are now unavailable or in an incorrect format. These articles must be removed from the dataset.

To identify these outliers, I fit a baseline logistic regression model and identified overly influential observations with Cook's Distance 4 times larger than the mean. I then manually inspected each of the 575 potential outliers and labeled true outliers, such as pages who's text is simply "404 Error: Page not found," or articles that are actually just lists. Out of the 575 potential outliers, I found 422 to be true outliers.

For further error removal, I intend to use a similar process to manually isolate articles including the text "error," and either too long or too short, and identify whether these are indeed outliers.

Group Differences

This section includes preliminary data exploration demonstrating differences in each variable accross the two labels. First, I incude tables with significance levels of a Mood's Median hypothesis test calculated using the RVAideMemoire package (Hervé 2019). Next, I include plots showing 95% confidence intervals of medians for each variable by group calculated using bootstrapping.

Mood's Median Test

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Complexity Metrics

Variable	Median - Fake	Median - Real	p-value
ARI	10.2	9.9	0.00
CL	48.6	49.0	0.00
FK	9.8	9.7	0.39
FOG	12.4	12.3	0.09
iqr_noun_phrase	0.5	1.0	0.00
iqr_sentence	5.0	5.0	0.00
iqr_verb_phrase	5.0	4.2	0.00
mu_noun_phrase	3.9	4.0	0.00
mu_sentence	12.4	12.1	0.00
mu_verb_phrase	7.6	7.5	0.00
num_verb_phrase	79.0	66.0	0.00
sd_noun_phrase	2.2	2.3	0.00
sd_sentence	4.1	4.0	0.01
sd_verb_phrase	3.6	3.5	0.00
SMOG	11.6	11.5	0.12
swc	23.0	21.5	0.00
tokens	460.0	435.0	0.00
TTR	0.5	0.5	0.49
types	228.0	218.0	0.00
wlen	4.0	4.0	0.00

LIWC Dictionary

- `summary`

Variable	Median - Fake	Median - Real	p-value
Analytic	83	88	0.00
Authentic	17	23	0.00
Clout	83	80	0.00
Dic	80	78	0.00
Sixltr	20	20	0.34
Tone	53	65	0.00
WC	372	346	0.00
WPS	19	19	0.65

- `function`

Variable	Median - Fake	Median - Real	p-value
adverb	3.97	3.35	0.00
article	7.25	7.27	0.74
auxverb	7.46	6.09	0.00
conj	5.42	5.17	0.00
function	46.89	44.44	0.00
i	0.22	0.65	0.00

Variable	Median - Fake	Median - Real	p-value
ipron	3.85	3.57	0.00
negate	1.15	0.81	0.00
ppron	6.52	6.39	0.09
prep	13.76	13.89	0.01
pronoun	10.50	10.07	0.00
shehe	3.41	2.62	0.00
they	0.76	0.44	0.00
we	0.37	0.33	0.05
you	0.28	0.40	0.00

- `othergram`

Variable	Median - Fake	Median - Real	p-value
adj	4.4	4.6	0.00
compare	2.3	2.3	0.66
interrog	1.1	1.1	0.05
number	2.2	2.4	0.00
quant	1.8	1.9	0.88
verb	14.0	12.6	0.00

- `affect`

Variable	Median - Fake	Median - Real	p-value
affect	4.40	4.58	0
anger	0.25	0.10	0
anx	0.12	0.00	0
negemo	1.22	0.95	0
posemo	2.76	3.13	0
sad	0.11	0.00	0

- `social`

Variable	Median - Fake	Median - Real	p-value
family	0.86	0.43	0
female	2.41	1.84	0
friend	0.30	0.20	0
male	1.53	1.01	0
social	13.38	11.65	0

- `cogproc`

Variable	Median - Fake	Median - Real	p-value
cause	1.35	1.09	0
certain	1.11	0.98	0
cogproc	9.01	7.59	0
differ	2.46	1.80	0
discrep	0.84	0.62	0
insight	1.67	1.54	0

Variable	Median - Fake	Median - Real	p-value
tentat	1.67	1.41	0

- `percept`

Variable	Median - Fake	Median - Real	p-value
feel	0.30	0.26	0
hear	0.79	0.71	0
percept	2.75	3.23	0
see	1.21	1.58	0

- `bio`

Variable	Median - Fake	Median - Real	p-value
bio	1.33	1.48	0.00
body	0.29	0.34	0.00
health	0.37	0.32	0.00
ingest	0.00	0.00	0.98
sexual	0.00	0.00	0.10

- `drives`

Variable	Median - Fake	Median - Real	p-value
achieve	1.09	1.26	0.00
affiliation	3.20	2.64	0.00
drives	7.87	7.55	0.00
power	2.29	2.37	0.00
reward	1.10	1.15	0.02
risk	0.24	0.12	0.00

- `timeorient`

Variable	Median - Fake	Median - Real	p-value
focusfuture	0.9	0.8	0
focuspast	4.5	4.2	0
focuspresent	8.0	6.9	0

- `relativ`

Variable	Median - Fake	Median - Real	p-value
motion	1.5	1.6	0.00
relativ	14.6	15.0	0.00
space	6.9	7.0	0.04
time	6.2	6.2	0.53

- `personc`

Variable	Median - Fake	Median - Real	p-value
death	0.00	0.00	0.00
home	0.28	0.26	0.03
leisure	1.81	2.05	0.00
money	0.25	0.20	0.00
relig	0.00	0.00	0.04
work	1.68	1.70	0.45

- `informal`

Variable	Median - Fake	Median - Real	p-value
assent	0.0	0.00	0.01
filler	0.0	0.00	0.11
informal	0.4	0.39	0.13
netspeak	0.0	0.00	0.00
nonflu	0.0	0.00	0.91
swear	0.0	0.00	0.49

- `punc`

Variable	Median - Fake	Median - Real	p-value
AllPunc	20.12	20.60	0.00
Apostro	2.79	2.55	0.00
Colon	0.23	0.35	0.00
Comma	5.39	5.40	0.93
Dash	0.87	0.92	0.03
Exclam	0.14	0.02	0.00
OtherP	0.35	0.43	0.00
Parenth	0.00	0.29	0.00
Period	4.92	5.10	0.00
QMark	0.06	0.00	0.00
Quote	2.88	2.34	0.00
SemiC	0.00	0.00	0.00

Parts-Of-Speech Tagging

Variable	Median - Fake	Median - Real	p-value
CC	12	11	0.00
CD	6	6	0.06
DT	32	30	0.00
EX	0	0	0.00
FW	0	0	0.00
IN	43	39	0.00
JJ	22	21	0.13
JJR	1	1	0.34
JJS	1	1	0.00
LS	0	0	0.00
MD	3	2	0.00
NN	51	51	0.80
NNP	41	41	0.65

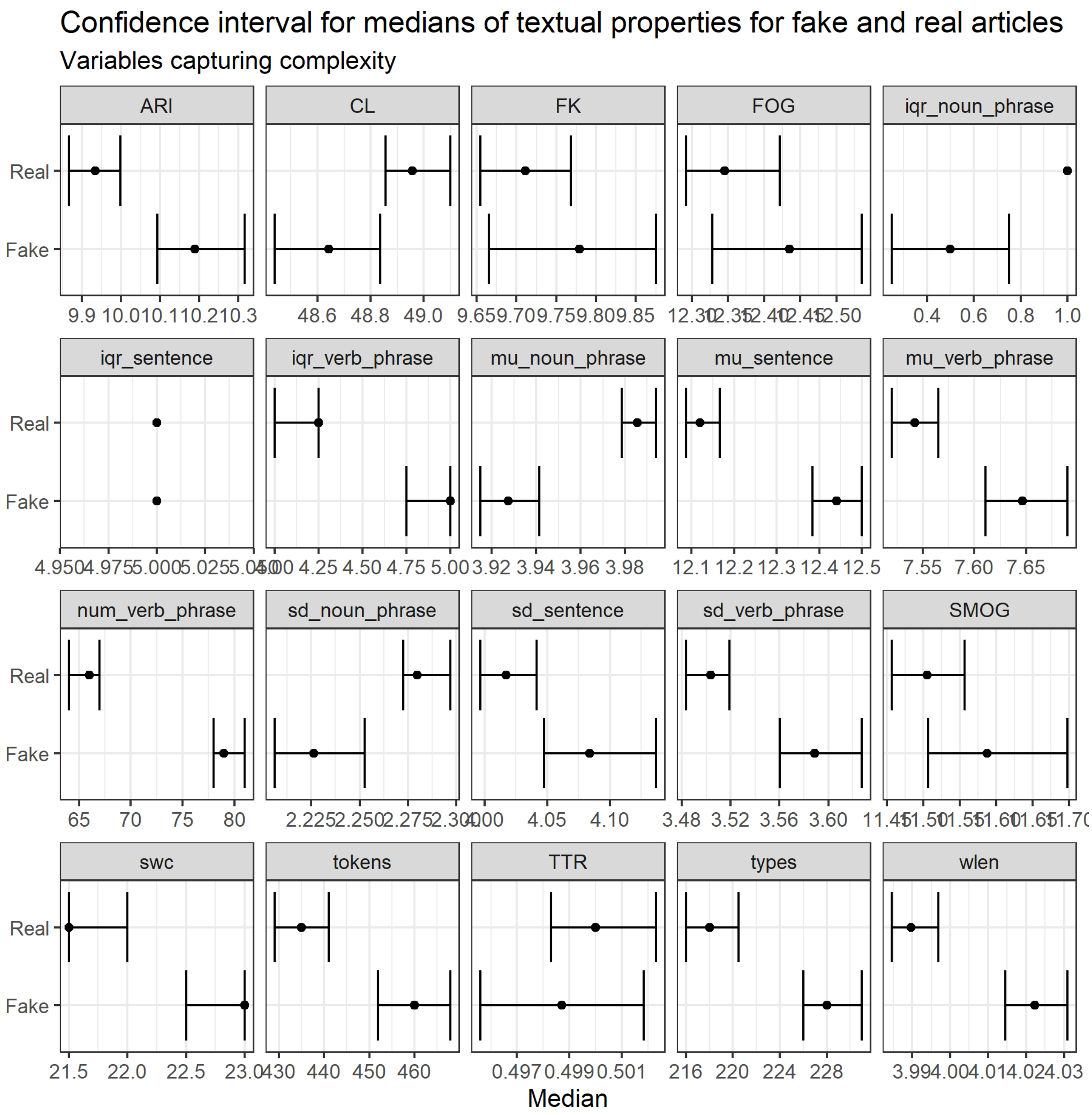
Variable	Median - Fake	Median - Real	p-value
NNPS	0	0	0.19
NNS	13	13	0.01
PDT	0	0	0.06
POS	3	3	0.00
PRP	18	16	0.00
PRP$	9	8	0.00
RB	20	15	0.00
RBR	0	0	0.95
RBS	0	0	0.00
RP	2	2	0.00
SYM	0	0	0.04
TO	10	8	0.00
UH	0	0	0.07
VB	13	11	0.00
VBD	15	14	0.02
VBG	10	8	0.00
VBN	8	7	0.00
VBP	6	6	0.08
VBZ	13	9	0.00
WDT	1	1	0.48
WP	1	1	0.11
WP$	0	0	0.02
WRB	2	1	0.00

Named Entity Recognition

Variable	Median - Fake	Median - Real	p-value
NER	56	50	0

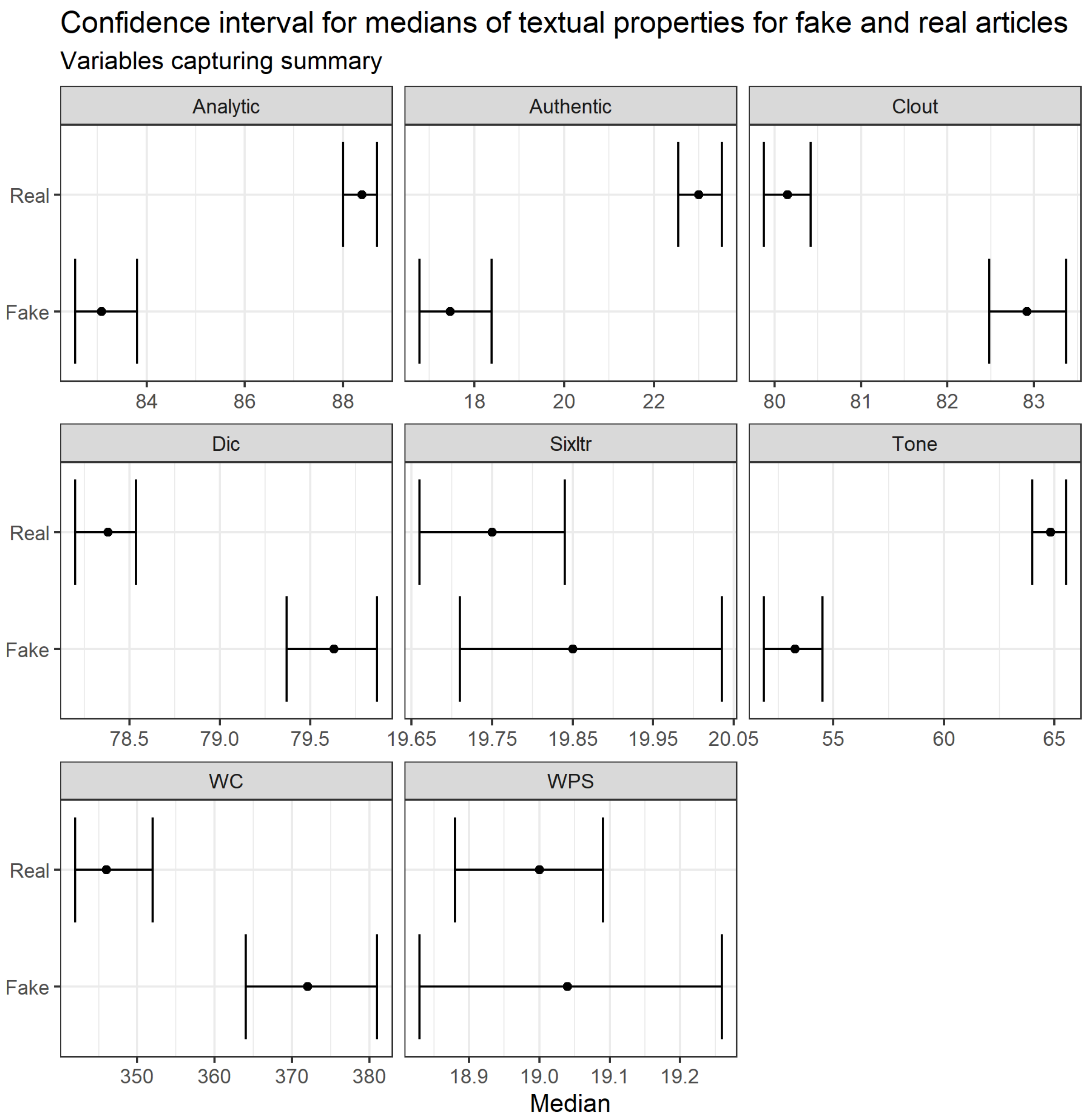
95% Median Confidence Intervals

Complexity Metrics

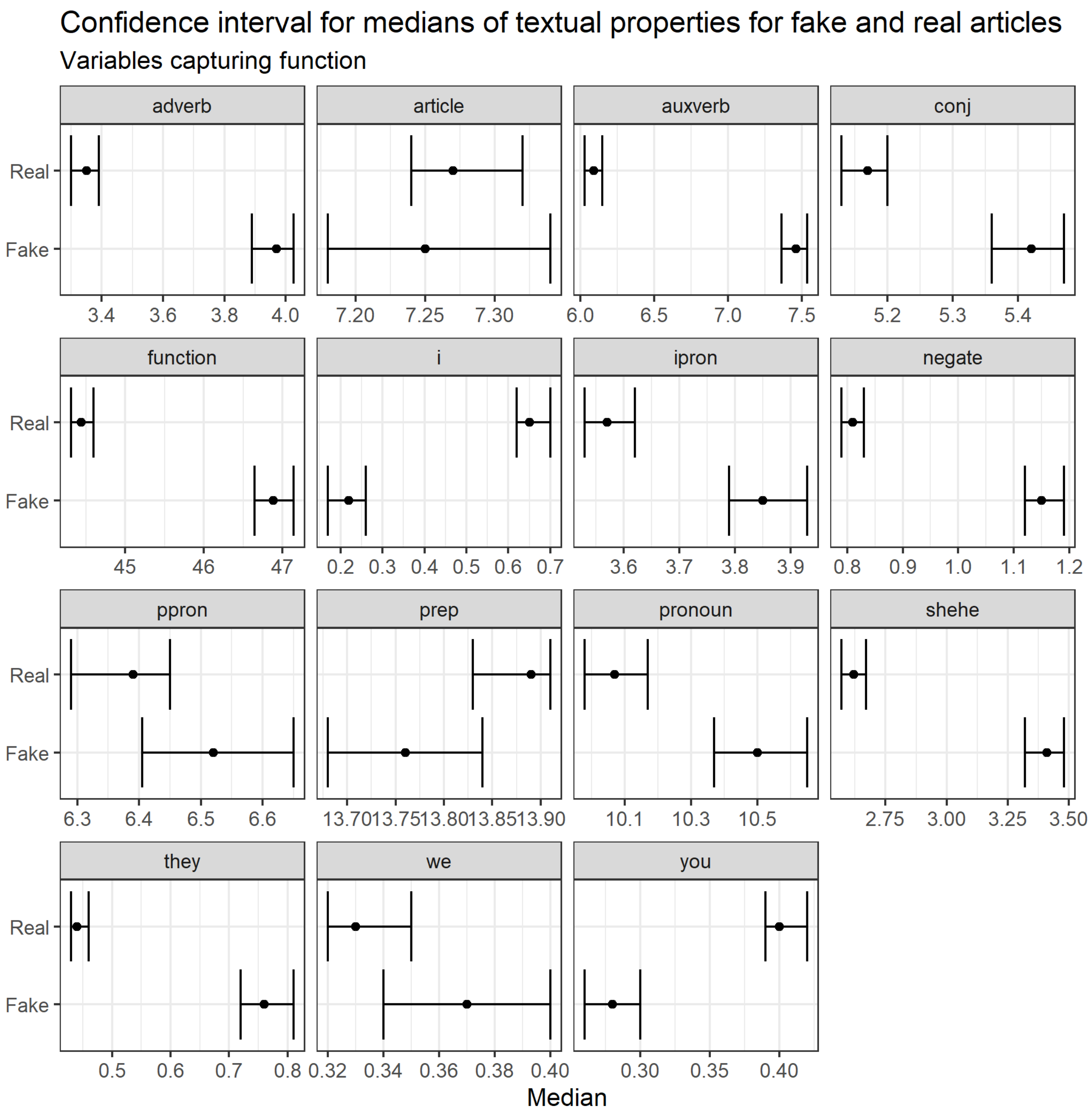


LIWC Dictionary

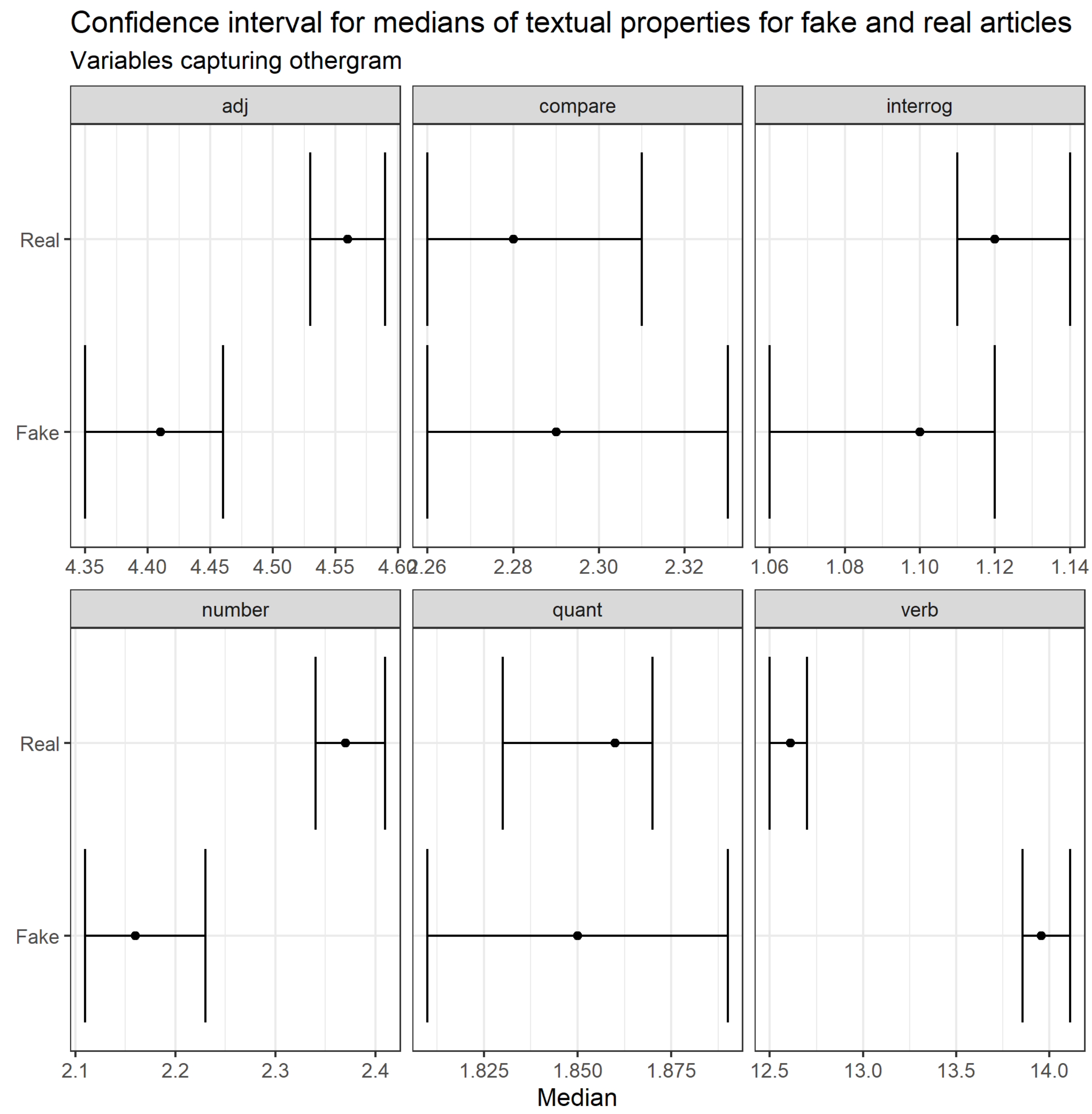
- `summary`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing summary
Analytic
Authentic
Clout
Dic
Sixltr
Tone
WC
WPS
Real
Fake
84
86
88
18
20
22
80
81
82
83
78.5
79.0
79.5
19.65
19.75
19.85
19.95
20.05
55
60
65
350
360
370
380
18.9
19.0
19.1
19.2
Median

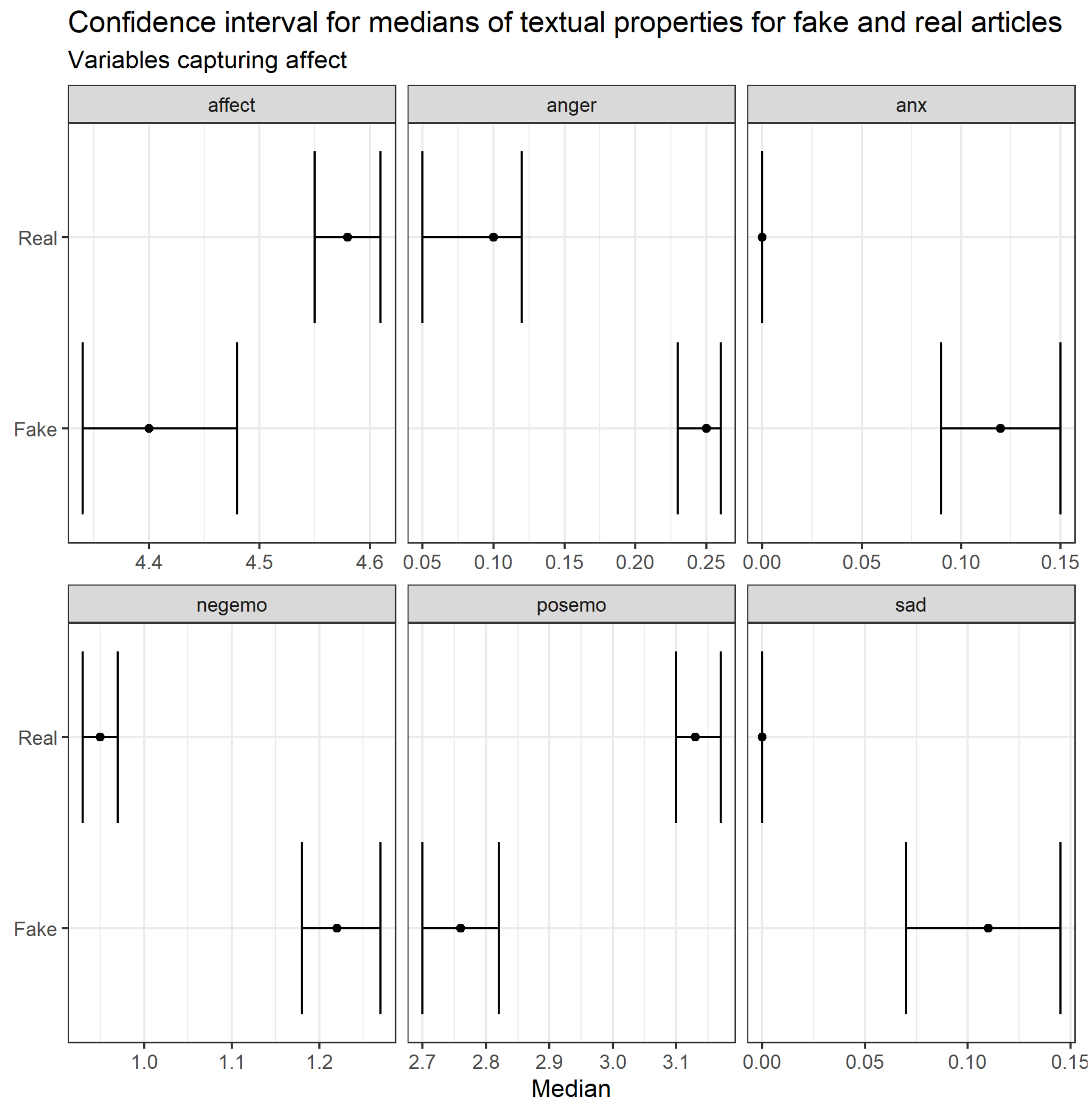
- `function`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing function
adverb
article
auxverb
conj
function
i
ipron
negate
ppron
prep
pronoun
shehe
they
we
you
Real
Fake
3.4
3.6
3.8
4.0
7.20
7.25
7.30
6.0
6.5
7.0
7.5
5.2
5.3
5.4
45
46
47
0.2
0.3
0.4
0.5
0.6
0.7
3.6
3.7
3.8
3.9
0.8
0.9
1.0
1.1
1.2
6.3
6.4
6.5
6.6
13.70
13.75
13.80
13.85
13.90
10.1
10.3
10.5
2.75
3.00
3.25
3.50
0.5
0.6
0.7
0.8
0.32
0.34
0.36
0.38
0.40
0.30
0.35
0.40
Median

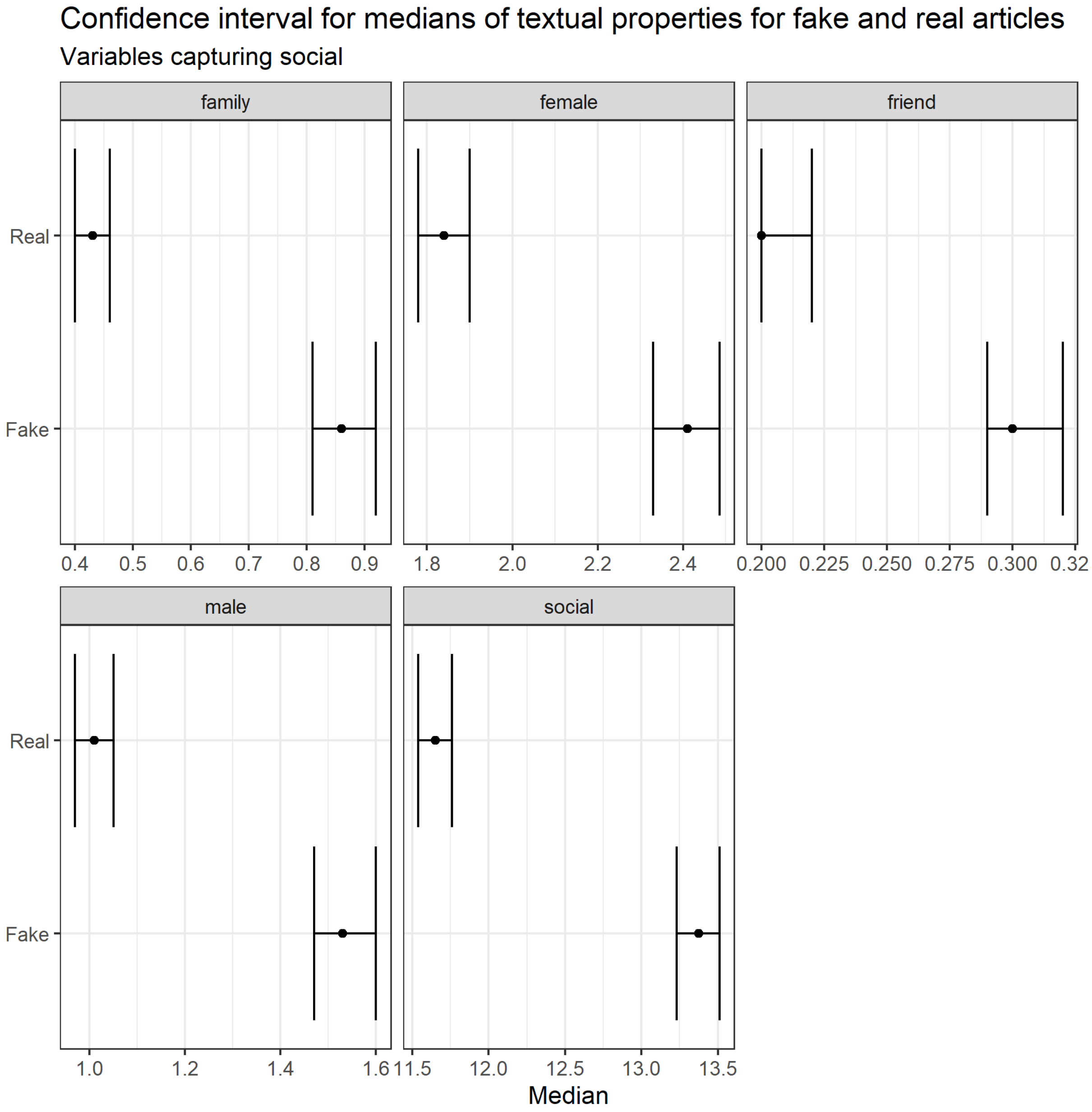
- `othergram`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing othergram
adj
compare
interrog
number
quant
verb
Real
Fake
Real
Fake
4.35
4.40
4.45
4.50
4.55
4.60
2.26
2.28
2.30
2.32
1.06
1.08
1.10
1.12
1.14
2.1
2.2
2.3
2.4
1.825
1.850
1.875
12.5
13.0
13.5
14.0
Median

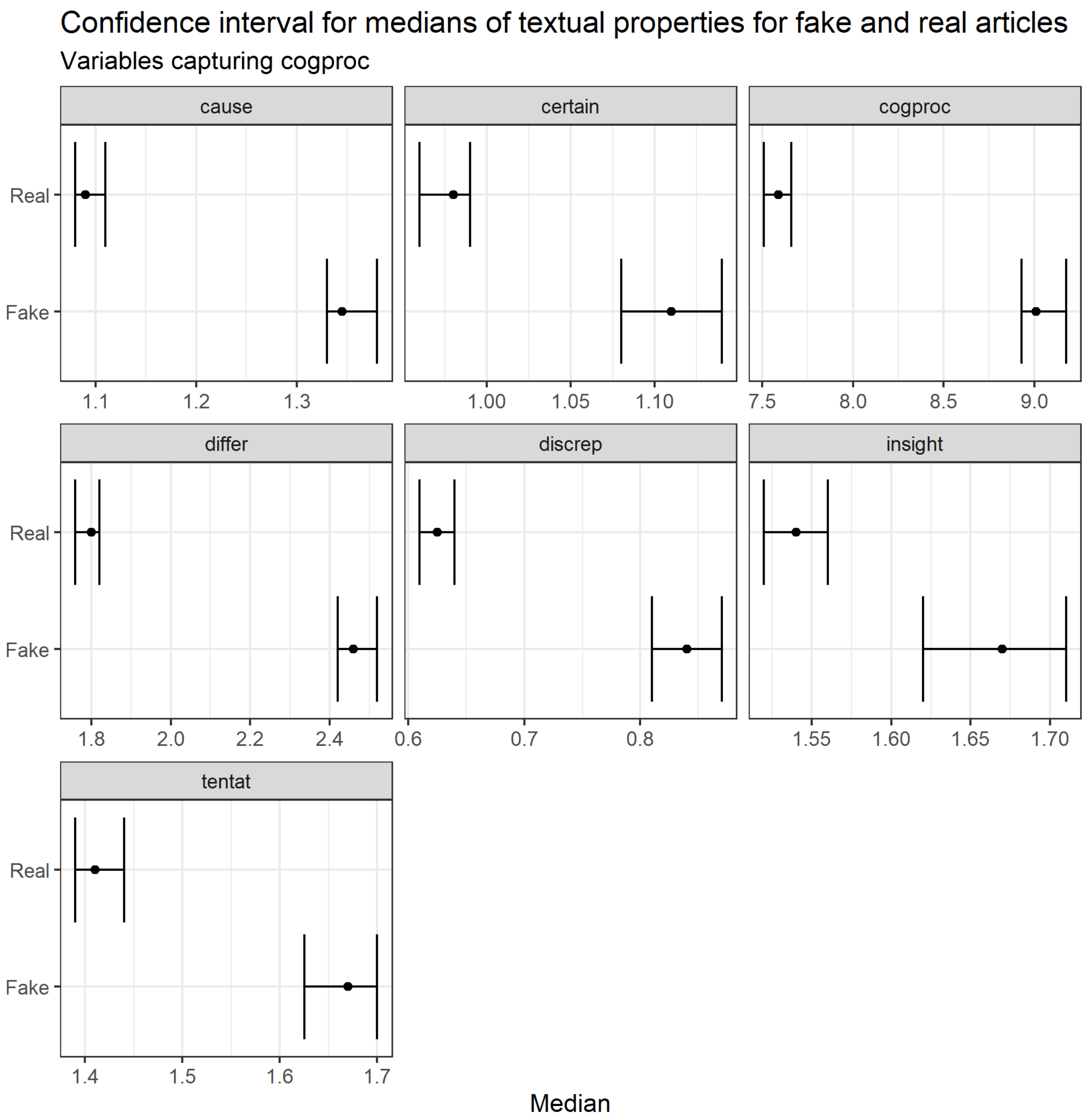
- `affect`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing affect
affect
anger
anx
negemo
posemo
sad
Real
Fake
4.4
4.5
4.6
0.05
0.10
0.15
0.20
0.25
0.00
0.05
0.10
0.15
1.0
1.1
1.2
2.7
2.8
2.9
3.0
3.1
0.00
0.05
0.10
0.15
Median

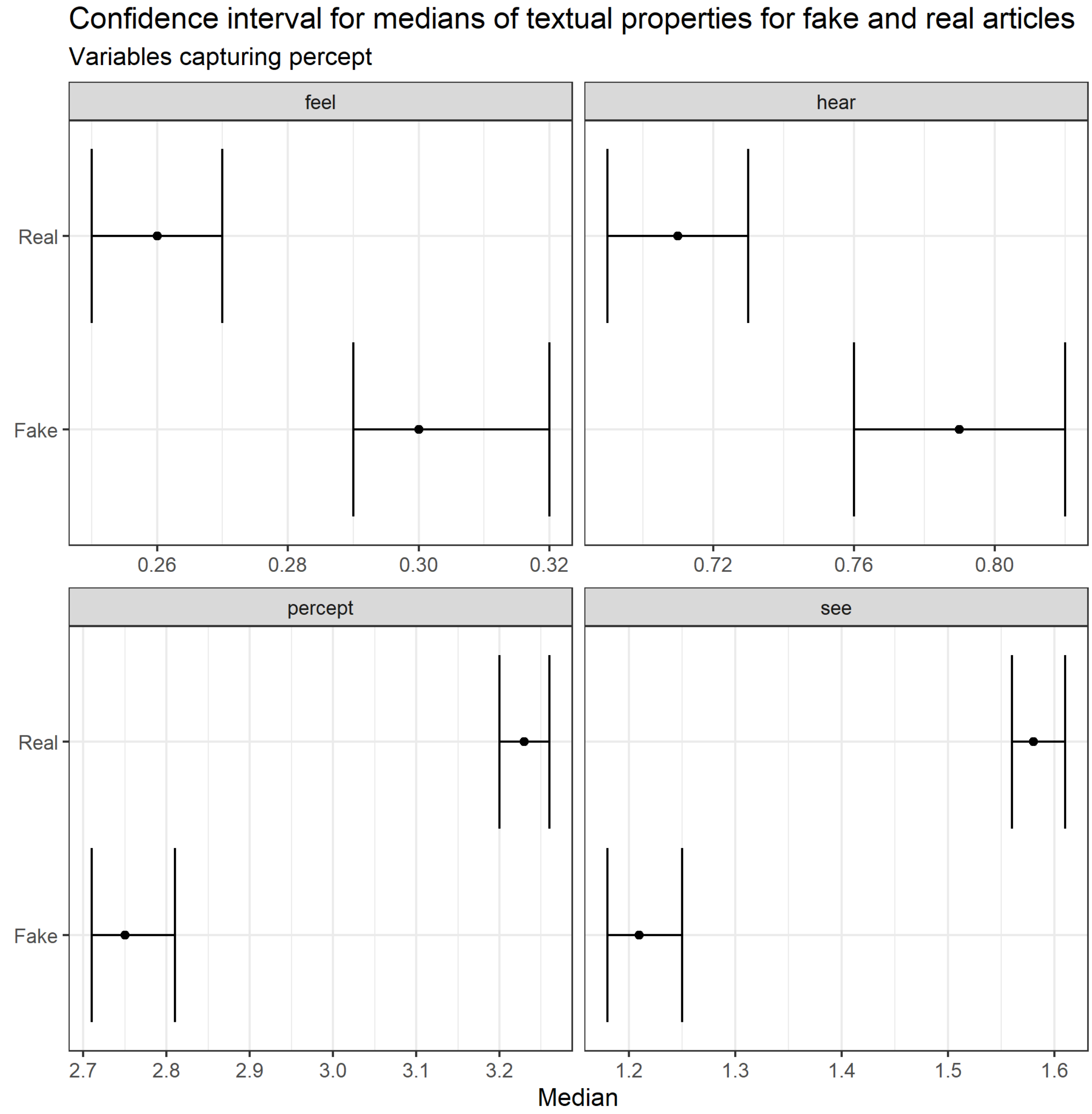
- `social`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing social
family
female
friend
male
social
Real
Fake
0.4
0.5
0.6
0.7
0.8
0.9
1.8
2.0
2.2
2.4
0.200
0.225
0.250
0.275
0.300
0.32
1.0
1.2
1.4
1.6
11.5
12.0
12.5
13.0
13.5
Median

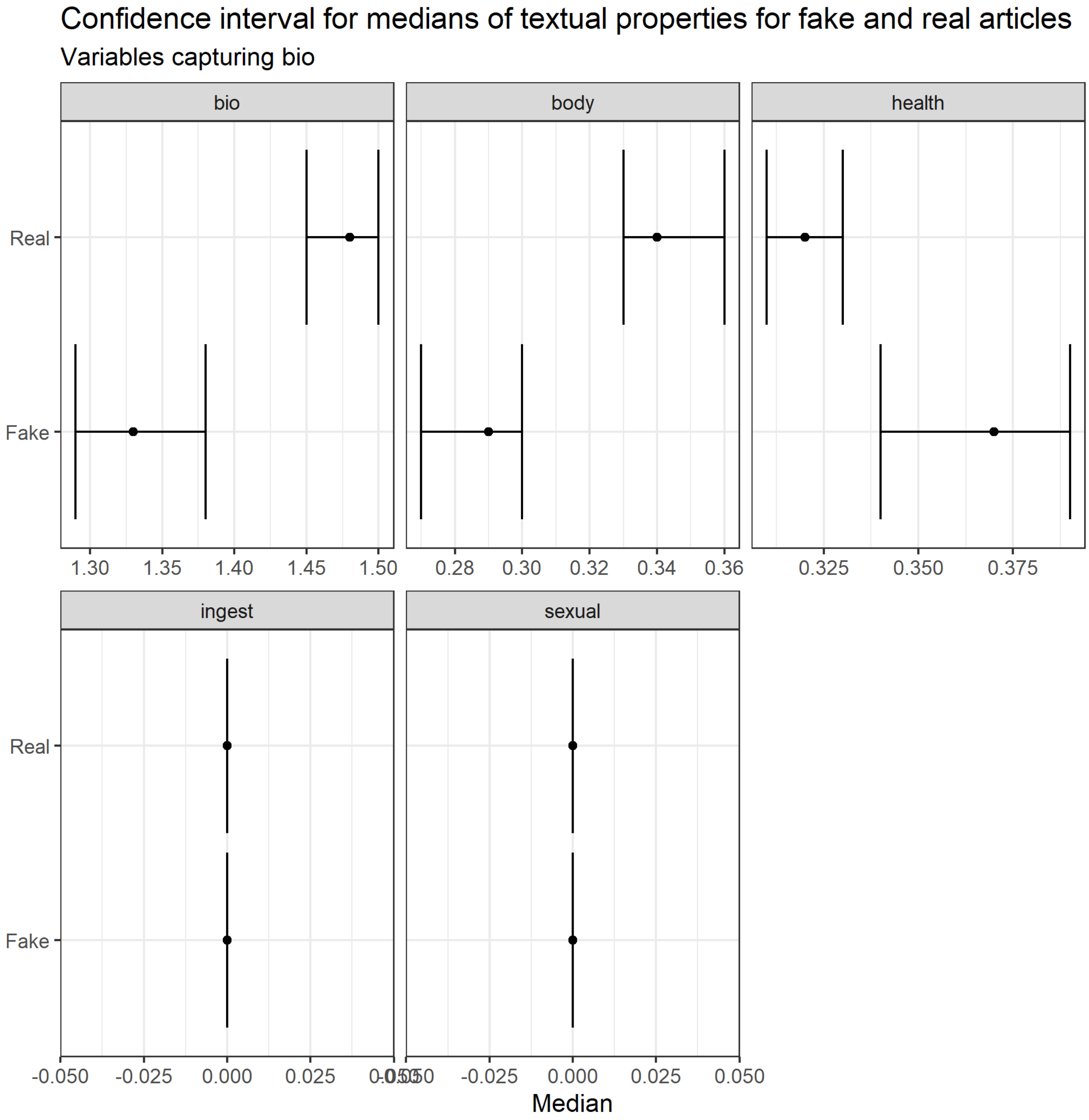
- `cogproc`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing cogproc
cause
certain
cogproc
differ
discrep
insight
tentat
Real
Fake
1.1
1.2
1.3
1.00
1.05
1.10
7.5
8.0
8.5
9.0
1.8
2.0
2.2
2.4
0.6
0.7
0.8
1.55
1.60
1.65
1.70
1.4
1.5
1.6
1.7
Median

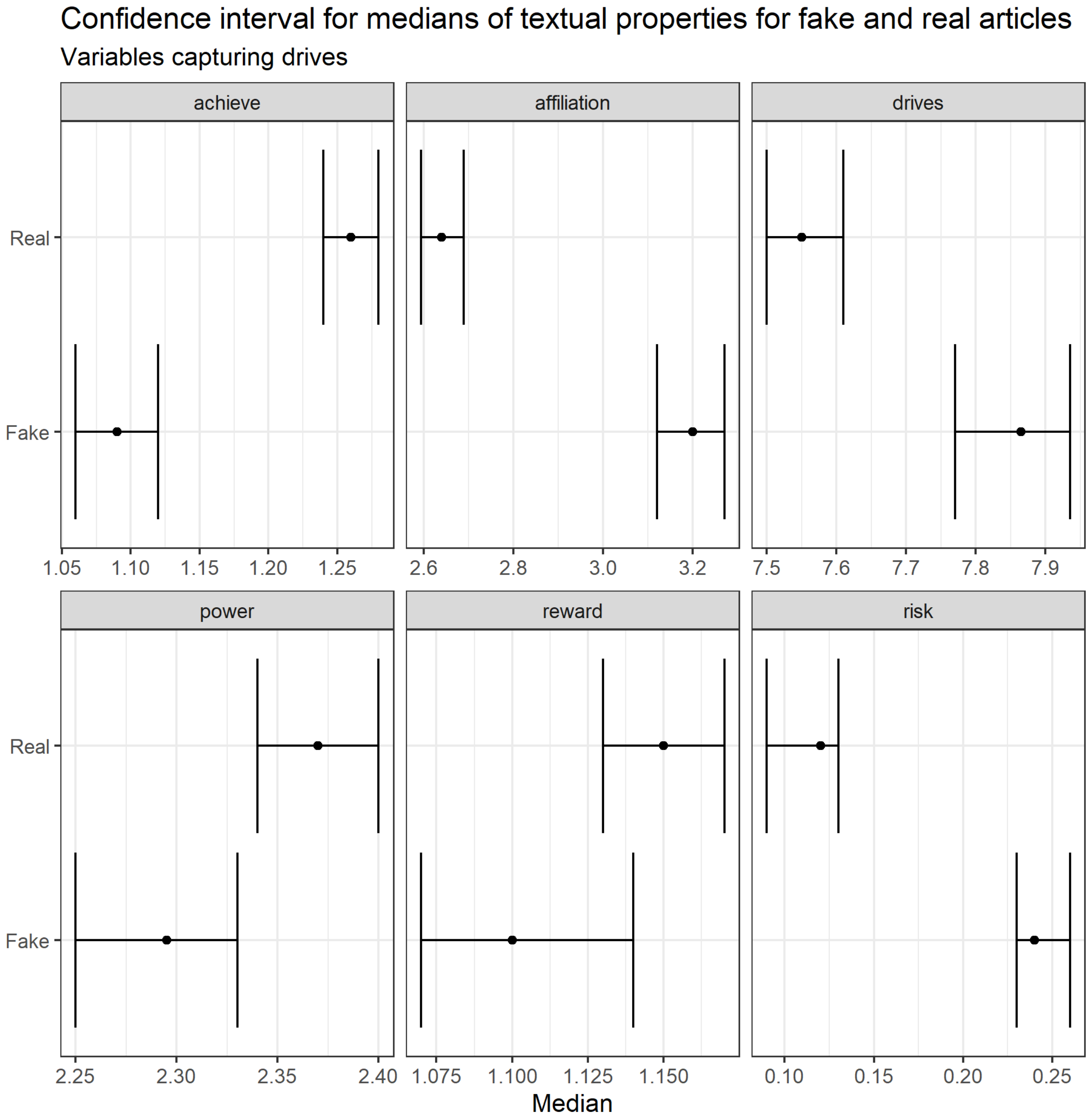
- `percept`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing percept
feel
hear
percept
see
Real
Fake
Real
Fake
0.26
0.28
0.30
0.32
0.72
0.76
0.80
2.7
2.8
2.9
3.0
3.1
3.2
1.2
1.3
1.4
1.5
1.6
Median

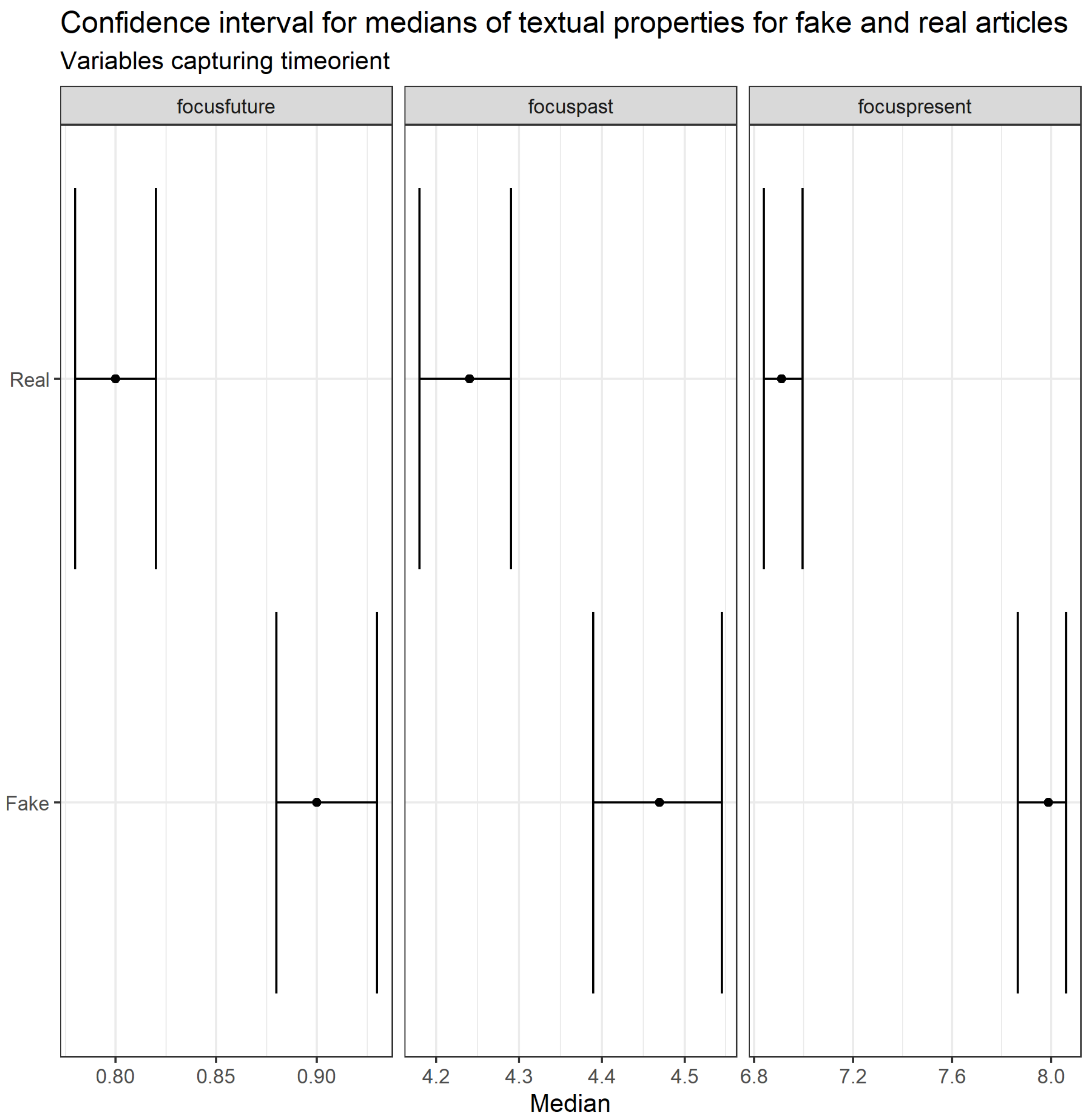
- bio


Confidence interval for medians of textual properties for fake and real articles
Variables capturing bio
bio
body
health
ingest
sexual
Real
Fake
Real
Fake
1.30
1.35
1.40
1.45
1.50
0.28
0.30
0.32
0.34
0.36
0.325
0.350
0.375
-0.050
-0.025
0.000
0.025
0.050
-0.050
-0.025
0.000
0.025
0.050
Median

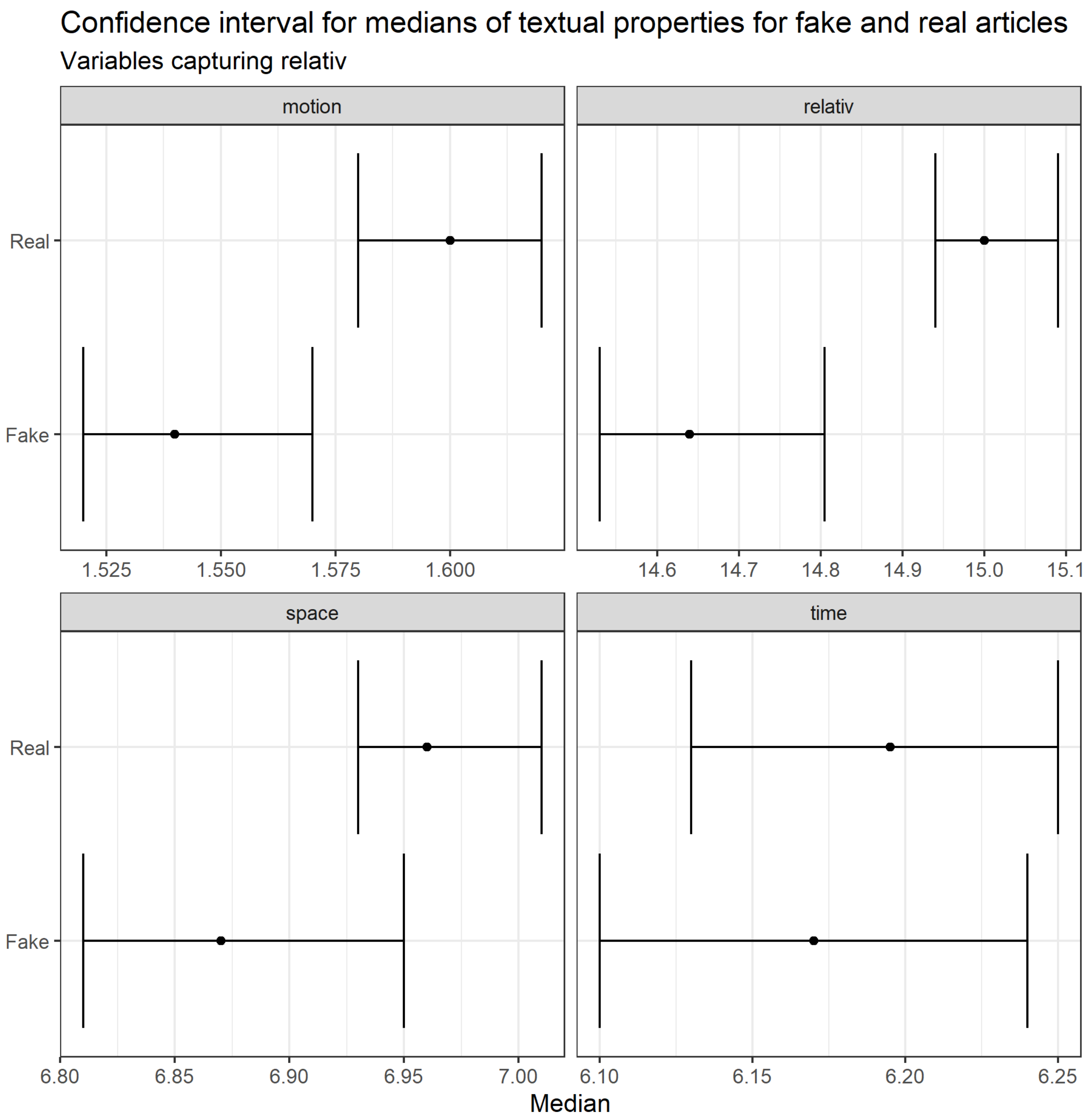
- `drives`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing drives
achieve
affiliation
drives
power
reward
risk
Real
Fake
Real
Fake
1.05
1.10
1.15
1.20
1.25
2.6
2.8
3.0
3.2
7.5
7.6
7.7
7.8
7.9
2.25
2.30
2.35
2.40
1.075
1.100
1.125
1.150
0.10
0.15
0.20
0.25
Median

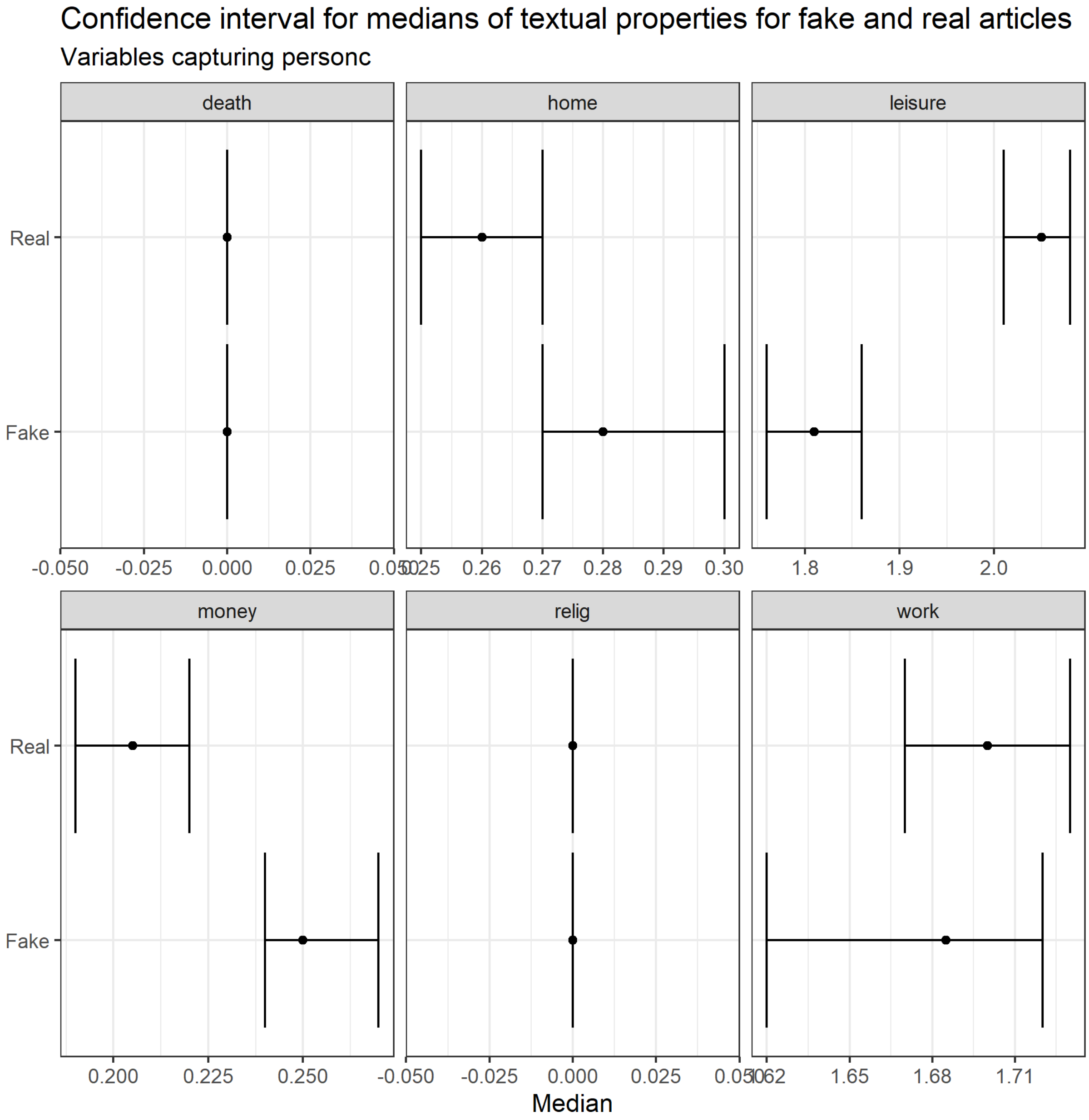
- `timeorient`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing timeorient
focusfuture
focuspast
focuspresent
Real
Fake
0.80
0.85
0.90
4.2
4.3
4.4
4.5
6.8
7.2
7.6
8.0
Median

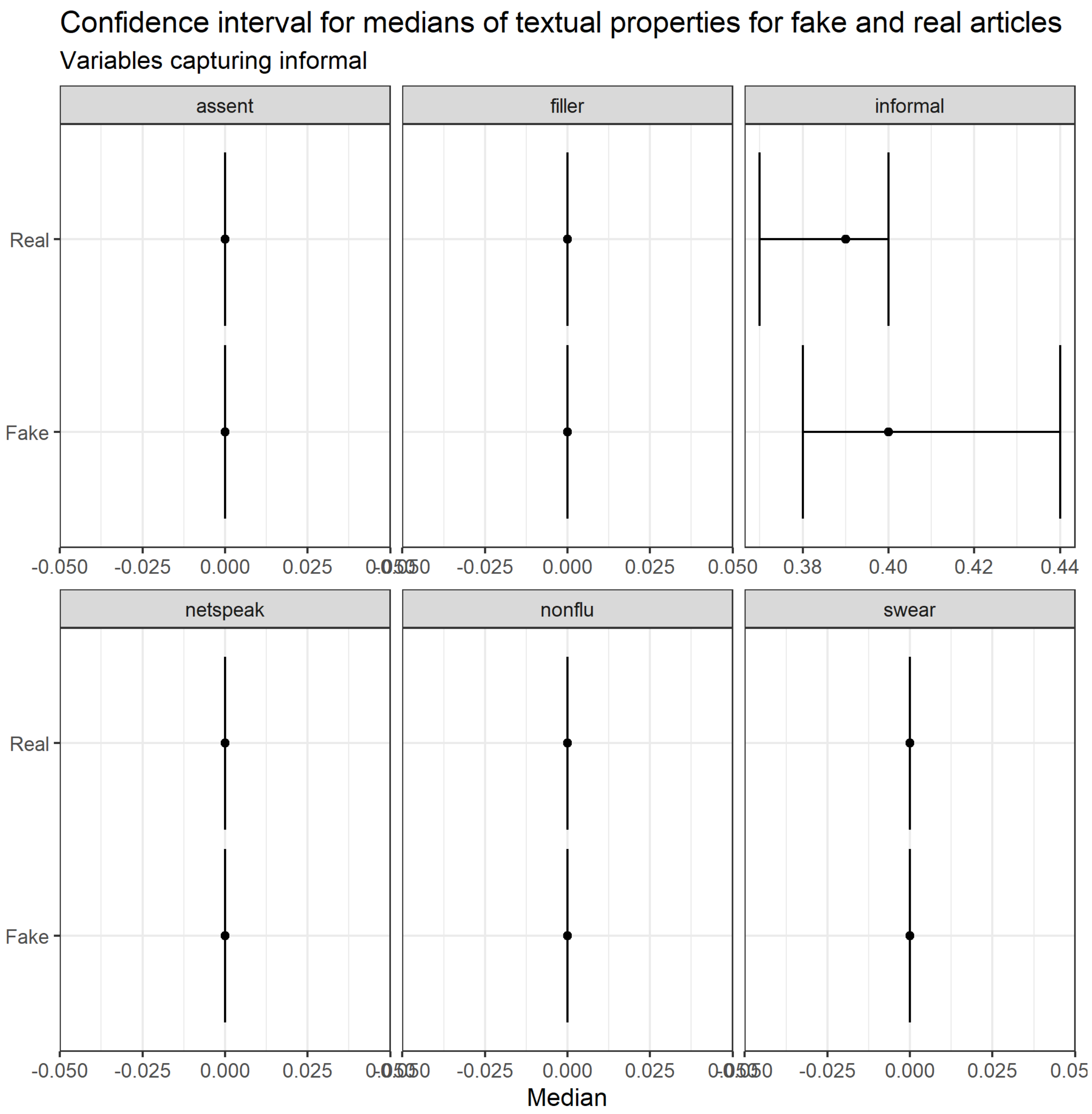
- `relativ`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing relativ
motion
relativ
space
time
Real
Fake
Real
Fake
1.525
1.550
1.575
1.600
14.6
14.7
14.8
14.9
15.0
15.1
6.80
6.85
6.90
6.95
7.00
6.10
6.15
6.20
6.25
Median

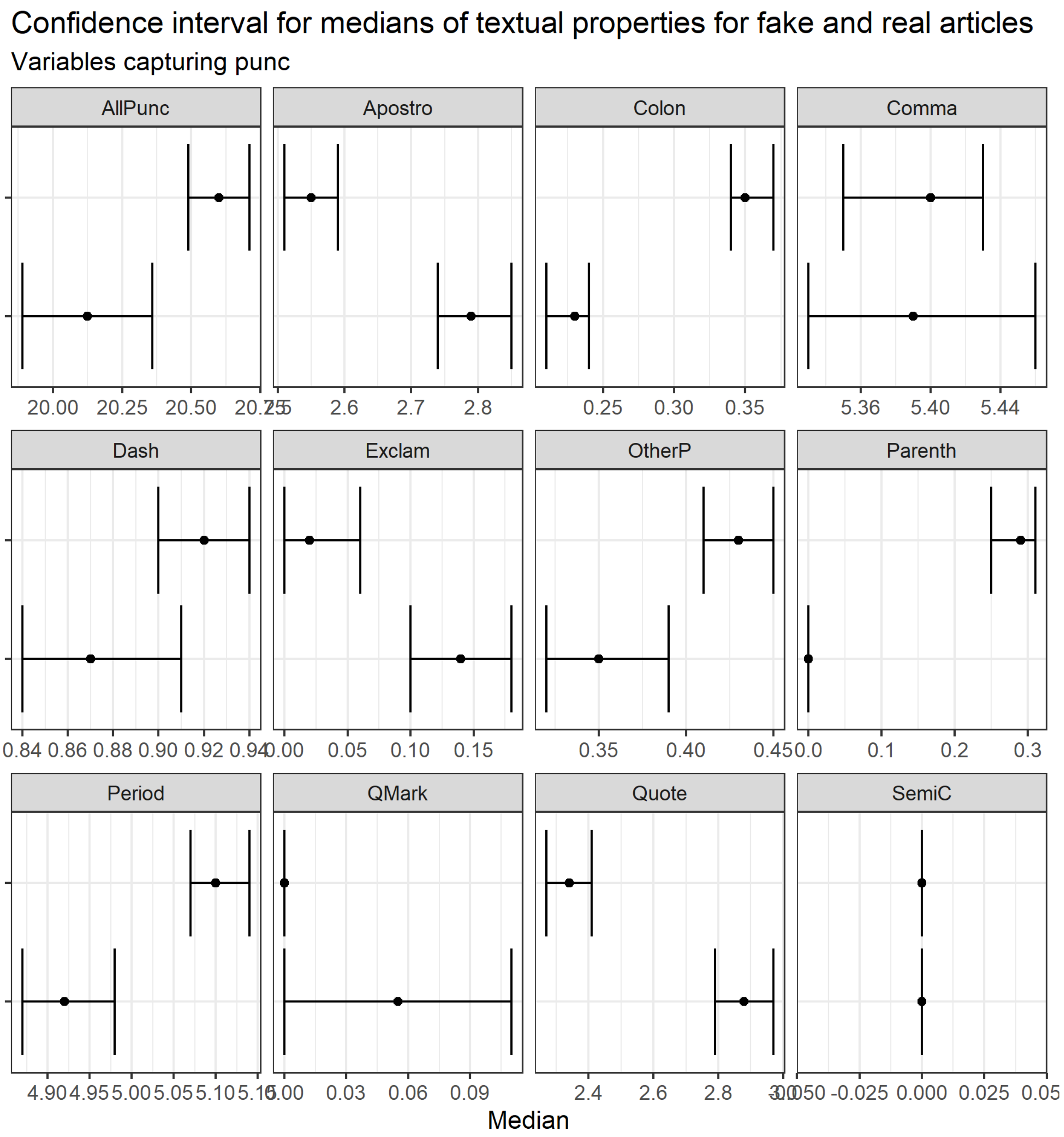
- `personc`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing personc
death
home
leisure
money
relig
work
Real
Fake
Real
Fake
-0.050
-0.025
0.000
0.025
0.050
0.25
0.26
0.27
0.28
0.29
0.30
1.8
1.9
2.0
0.200
0.225
0.250
-0.050
-0.025
0.000
0.025
0.050
1.62
1.65
1.68
1.71
Median

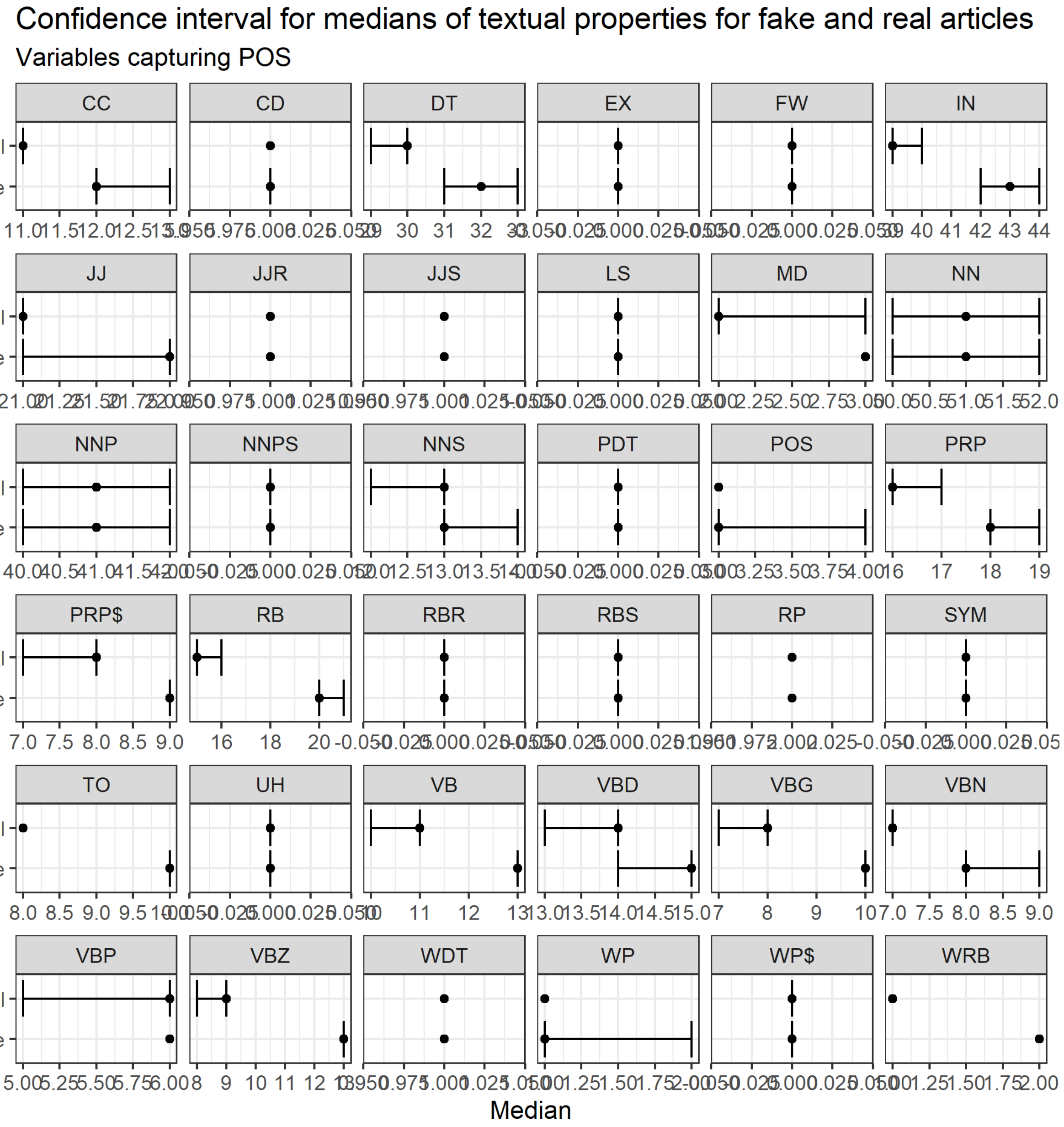
- `informal`


Confidence interval for medians of textual properties for fake and real articles
Variables capturing informal
assent
filler
informal
netspeak
nonflu
swear
Real
Fake
Real
Fake
-0.050
-0.025
0.000
0.025
0.050
0.38
0.40
0.42
0.44
0.05
Median

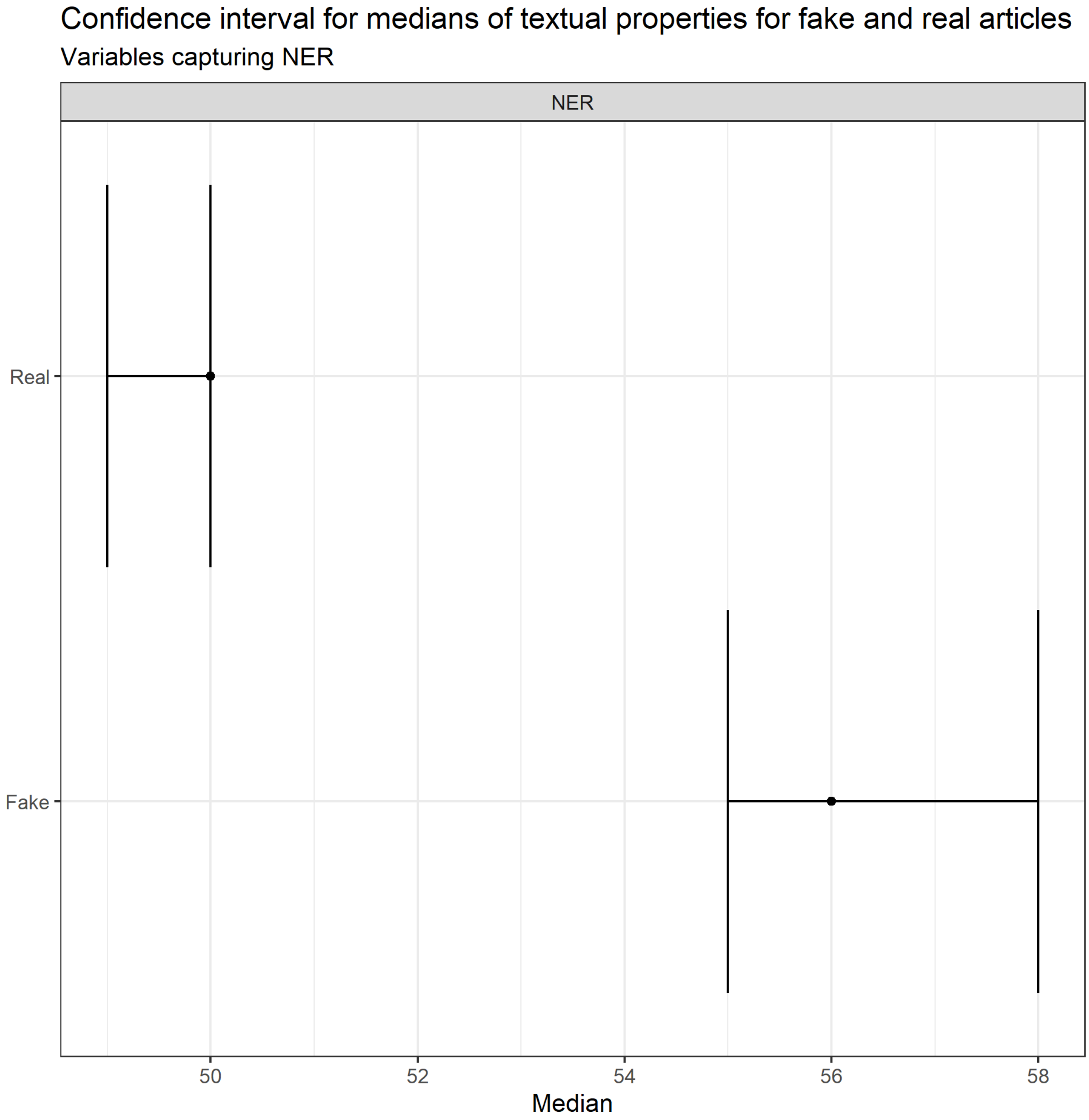
- `punc`



Parts-Of-Speech Tagging



Named Entity Recognition



Confidence interval for medians of textual properties for fake and real articles

Variables capturing NER

References

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. “Quanteda: An R Package for the Quantitative Analysis of Textual Data.” *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Hervé, Maxime. 2019. *RVAideMemoire: Testing and Plotting Procedures for Biostatistics*. https://CRAN.R-project.org/package=RVAideMemoire.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. “The Stanford CoreNLP Natural Language Processing Toolkit,” 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Pennebaker, James W., Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. “Linguistic Inquiry and Word Count: LIWC 2015.” *Pennebaker Conglomerates*. http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_OperatorManual.pdf.

Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. “FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media.” *CoRR* abs/1809.01286. http://arxiv.org/abs/1809.01286.